

**ANALISE *IN SILICO* E *IN VIVO* DA DIVERSIDADE NUCLEOTÍDICA EM *Coffea* spp.<sup>1</sup>**

Ramon Vidal<sup>2</sup>; Karina Yanagui<sup>3</sup>; Lucia Pires Ferreira<sup>4</sup>; Sergio Dias Lannes<sup>4</sup>; Luiz G. E. Vieira<sup>5</sup>; Jorge Mondego<sup>2</sup>; Marcelo F. Carazzolle<sup>2</sup>; Gonçalves A. Pereira<sup>2</sup>; David Pot<sup>6</sup>; Luiz Filipe P. Pereira<sup>7</sup>

<sup>1</sup> Trabalho financiado pelo Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café – CBP&D/Café

<sup>2</sup> LGE - UNICAMP, Campinas (SP) Brasil

<sup>3</sup> Bolsista Iniciação Científica – ProICI – LBI LBI/IAPAR CP 481 86001-970 Londrina (PR) Brasil

<sup>4</sup> Bolsista do CBP&D/Café IAPAR LBI-AMG CP 481 86001-970 Londrina (PR) Brasil

<sup>5</sup> CIRAD, UMR DAP, Montpellier, F-34398 France

<sup>6</sup> LBI-AMG IAPAR CP 481 86001-970 Londrina (PR) Brasil

<sup>7</sup> Embrapa Café – LBI/IAPAR CP 481 86001-970 Londrina (PR) Brasil lpereira@iapar.br

**RESUMO:** Polimorfismos de modificações nucleotídicas (SNPs – *Single Nucleotide Polymorphisms*, INDELs – *Insertion / Deletions*) têm uma alta frequência nos genomas da maioria dos organismos, incluindo plantas. Eles vêm se tornando a escolha principal de marcador para trabalhos de melhoramento, genotipagem e diagnóstico. A identificação destes polimorfismos irá fornecer marcadores que poderão ser utilizados para o mapeamento genético, estudos de genética de população e de associação. Portanto, os objetivos deste trabalho foram: 1) identificar *in silico* SNPs e INDELs existentes em seqüências de ESTs disponíveis; e 2) analisar a diversidade nucleotídica em *Coffea* spp. Um pipeline para identificação de SNPs e INDELs foi desenvolvido utilizando seqüências de ESTs disponíveis de *Coffea* spp. Foi utilizado uma estratégia para detecção de SNPs em dentro de 23.019 contigs. Um total 23.062 SNPs e 2.165 INDELs foram encontrados em 5184 contigs que continham pelo menos quatro ESTs. Análises *in silico* permitiram a identificação de diferentes alelos de *C. canephora* e *C. eugenioides* que estão presentes em *C. arabica*. A maioria dos ESTs de *C. arabica* vieram de apenas dois alelos, uma evidência molecular sobre a especiação de *C. arabica*. De acordo com essas análises cerca de 55% das seqüências de *C. arabica* são derivadas do genoma de *C. eugenioides* e 45% de *C. canephora*. Além disso, foi possível observar que o genoma de *C. eugenioides* contribui principalmente para genes relacionados a metabolismo basal, enquanto que os genes de *C. canephora* estão envolvidos com sinais de tradução e regulação da expressão gênica. Análises *in vivo* estão sendo realizadas através do sequenciamento de diversos genes em 24 genótipos de *Coffea* sendo 12 de *C. arabica*, 9 de *C. canephora* e três de outras espécies de *Coffea*, para uma análise maior da diversidade nucleotídica do gênero. Resultados referentes ao sequenciamento do gene de sacarose fosfato sintase (SPS) apresentaram 21 polimorfismos, sendo a maioria interespecíficos (*C. arabica*, *C. canephora*, *C. eugenioides* e *C. racemosa*). Para os genótipos de *C. canephora* foram observados nove polimorfismos intraespecíficos. Já os polimorfismos encontrados entre os genótipos de *C. arabica* forma os mesmos detectados entre *C. canephora* e *C. eugenioides*.

**Palavras-Chave:** SNP, INDEL, mapeamento, genoma

**IN SILICO AND IN VIVO ANALYSIS OF THE NUCLEOTIDE DIVERSITY IN *Coffea* spp.**

**ABSTRACT:** Single nucleotide polymorphisms are the most abundant polymorphisms in the genomes analyzed to date. They are becoming the main choice of molecular markers for breeding, genotyping, and diagnosis purposes, due to the large amount of sequences data available. Identification of those nucleotide polymorphisms will provide useful markers for genetic mapping, population genetics and association studies. It will also provide criteria to infer the evolutionary history of the analyzed genes, which can be relevant to select the best candidate genes to test in future association studies. For those reasons, the objectives of this work were: 1) identify and validate both *in silico* and *in vivo*, the SNPs and INDELs existing in EST resources; and 2) analyze the nucleotide diversity in *Coffea* spp., in addition to selected *C. arabica* cultivars. A Pipeline for identification of SNPs and INDELs was generated using sequences from the Brazilian ESTs Coffee Genome Project as well as other *Coffea* sequences available in GenBank. The pipeline was carried out by a haplotype –based strategy to detect reliable SNPs in 23.019 contigs assembled. A total of 23.062 SNPs e 2.165 INDELs were identified in 5184 contigs with more than four ESTs assembled. With the haplotype-based strategy, it was possible to define the probable ancestral of *C. arabica* transcripts. The majority of ESTs from *C. arabica*, came from only two different alleles, providing molecular evidences about *C. arabica* speciation. According to our analysis, approximately 55% of *C. arabica* sequences were derived from *C. eugenioides*, and 45% were considered as come from *C. canephora*. Interestingly, *C. eugenioides* contributes mostly with genes related to basal metabolism and the secondary metabolism, while genes *C. canephora* genes are involved with signal transduction and gene expression regulation. The *in vivo* analyses are being performed by sequencing PCR fragments of several genes in 24 *Coffea* genotypes corresponding respectively to 12 *C. arabica*, 9 *C. canephora* and 3 *Coffea* spp. Genotypes belonging to *C. arabica* and *C. canephora* were chosen in order to represent the largest diversity possible. Sequencing results from Sucrose Phosphate Sintase gene in those genotypes reveal the presence of SNPs mainly in interspecific sequences. A higher number of SNPs intraspecific was also observed for *C. canephora*. Intraespecific SNPs for *C. arabica* were the same observed in the two ancestral genomes, *C. canephora* and *C. eugenioides*.

**Key words: SNP, INDEL, mapping, genome**

## INTRODUÇÃO

A utilização de ferramentas biotecnológicas, como marcadores moleculares pode acelerar programas de melhoramento através de seleção assistida por marcadores. No entanto, para *C. arabica*, o número de marcadores polimórficos disponíveis é pequeno, comparado com outras importantes espécies vegetais, dificultando o desenvolvimento deste tipo de estratégia. Uma das razões para isso é a baixa diversidade existente dentro dos acessos de *C. arabica*, dificultando o trabalho de busca de marcadores polimórficos. Polimorfismos de modificações nucleotídicas (SNPs – *Single Nucleotide Polymorphisms*, INDELs – *Insertion / Deletions*) têm uma alta frequência nos genomas da maioria dos organismos, incluindo plantas. Eles vem se tornando a escolha principal de marcador para trabalhos de melhoramento, genotipagem e diagnóstico. Os SNPs e INDELs podem ser detectados através de várias técnicas, a maioria delas envolvendo um conhecimento prévio das seqüências de DNA de diferentes genótipos, o que no passado encarecia o custo de implementação para a sua identificação. Atualmente, projetos de sequenciamento em larga escala oferecem a possibilidade da descoberta de polimorfismos de nucleotídeos a custos mais baixos, pois variações podem ser encontradas computacionalmente pela análise da sua redundância em banco de dados.

Com os dados de ESTs disponíveis, tanto do projeto Genoma Café, como de bancos públicos, foi desenvolvido um pipeline de identificação de polimorfismos de nucleotídeos (LGE – <http://www.lge.ibi.unicamp.br/cafe/>). Neste trabalho reportamos a validação *in silico* de SNPs e INDELs a partir deste pipeline, assim com um análise da diversidade através da análise de seqüências de uma população maior de genótipos e espécies de *Coffea* sequenciadas previamente.

## MATERIAL E MÉTODOS

Para a montagem híbrida entre os transcriptomas de *Coffea arabica* e *Coffea canephora* um total de 267533 ESTs, 78,182 de *C. canephora* e 189,351 de *C. arabica*, sendo 6 genótipos de *canephora* e apenas 2 de *arabica* (Mundo Novo e Catuai). Esses dados são derivados de 48 bibliotecas que foram coletadas do Projeto Genoma Café Brasileiro e de seqüências disponibilizadas em banco públicos (Lin et al 2006).

Antes da montagem as seqüências foram trimadas, removendo seqüência de vetor, seqüências ribossomais, caudas polyA/T e descartando seqüências de baixa qualidade. Toda a montagem foi feita com o programa CAP3, cujos parâmetros foram ajustados para minimizar a ocorrência de agrupamentos de parálogos, permitindo uma sobreposição mínima de 100 bp com similaridade mínima de 95%.

QualitySNP (<http://www.bioinformatics.nl/tools/snpweb/>) foi o programa utilizado para descoberta de SNPs *in silico*. Este software utiliza uma estratégia baseada na reconstrução de haplótipos baseado nos padrões de SNPs encontrados nos ESTs utilizados na montagem dos Contigs híbridos de *C. arabica* e *C. canephora*. Para filtrar falsos SNPs o QualitySNP faz um filtro pela qualidade do cromatograma e pela redundância de seqüências. Todas as informações geradas pelo QualitySNP sobre os Contigs, ESTs e SNPs (incluindo informações dos haplótipos, posição dos SNPs, etc) são armazenadas em um bando de dados mysql. Essa base de dados também possui informações sobre anotação automática e manual dos genes. Scripts em PERL foram desenvolvidos para minerar esse banco de dados e buscar similaridades entre haplótipos de *C. arabica* e *C. canephora*. Os ESTs de haplótipos de *arabica* que agruparam com ESTs de haplótipos de *canephora* foram considerados como ESTs de origem do ancestral *C. canephora* (CAcc) e, por exclusão, os que não foram similares com *canephora* foram considerados ESTs de origem do ancestral *C. eugenioides* (CAce).

Para os trabalhos de validação dos SNPs *in vivo* e análise de diversidade nucleotídica foram coletadas folhas dos seguintes genótipos: - oito de *Coffea arabica* da coleção da Etiópia do IAPAR(E007, E237, E017, E238, E123A, E123B, E464, E516), representativos do centro de origem da *C. arábica*, e genótipos comerciais de *Coffea arabica*: cv Typica e cv Bourbon, Catuai e cv Mundo Novo. Também foram coletadas folhas de *Coffea canephora* G21. *Coffea eugenioides* *Coffea racemosa* e *Psilanthus bengalensis*. A extração do DNA foi realizada seguindo o protocolo do método MATAB 2% de acordo com protocolo de Risterucci *et al.* (2000) com modificações.

Foram realizadas PCR com as amostras de DNA acima com *primers* que amplificam regiões de genes que codificam para açúcares (invertase de parede celular, invertase vacuolar, fosfato sacarose sintase (SPS), sacarose sintase 1, Sacarose sintase 2,) diterpenos (kaureno oxidase, copalil difosfato sintase, kaureno sintase) Álcool desidrogenase (ADH1.1), DNA cloroplástico (rpl16) e mitocondrial (Cox 3).

Os fragmentos de DNA amplificados foram purificados e quantificados para sequenciamento dos genes utilizando o ABI3130 XL Genetic analyzer. Os produtos de PCR foram seqüenciados e analisados utilizando o software Códon Code Aligner.

## RESULTADOS E DISCUSSÃO

Através da construção do pipeline (Figura 1), um total 23.062 SNPs e 2.165 INDELS foram encontrados em 5184 contigs que continham pelo menos quatro ESTs. Os polimorfismos detectados aos níveis interspecíficos (*Coffea arabica* vs *Coffea canephora*) dentro do banco de dados EST (busca in silico) foram verificados. Observou-se a ocorrência de dois grupos de haplótipos dentro da espécie aloploidica *Coffea arabica*, cada um apresentando alta similaridade com *Coffea canephora* e *Coffea eugenoides* respectivamente. É importante notar que a maioria dos polimorfismos encontrados dentro de *Coffea arabica* está no estado heterozigoto e que esses polimorfismos correspondem aos alelos detectados respectivamente dentro de *C. canephora* e *C. eugenoides* (espécies ancestrais de *C. arabica*).

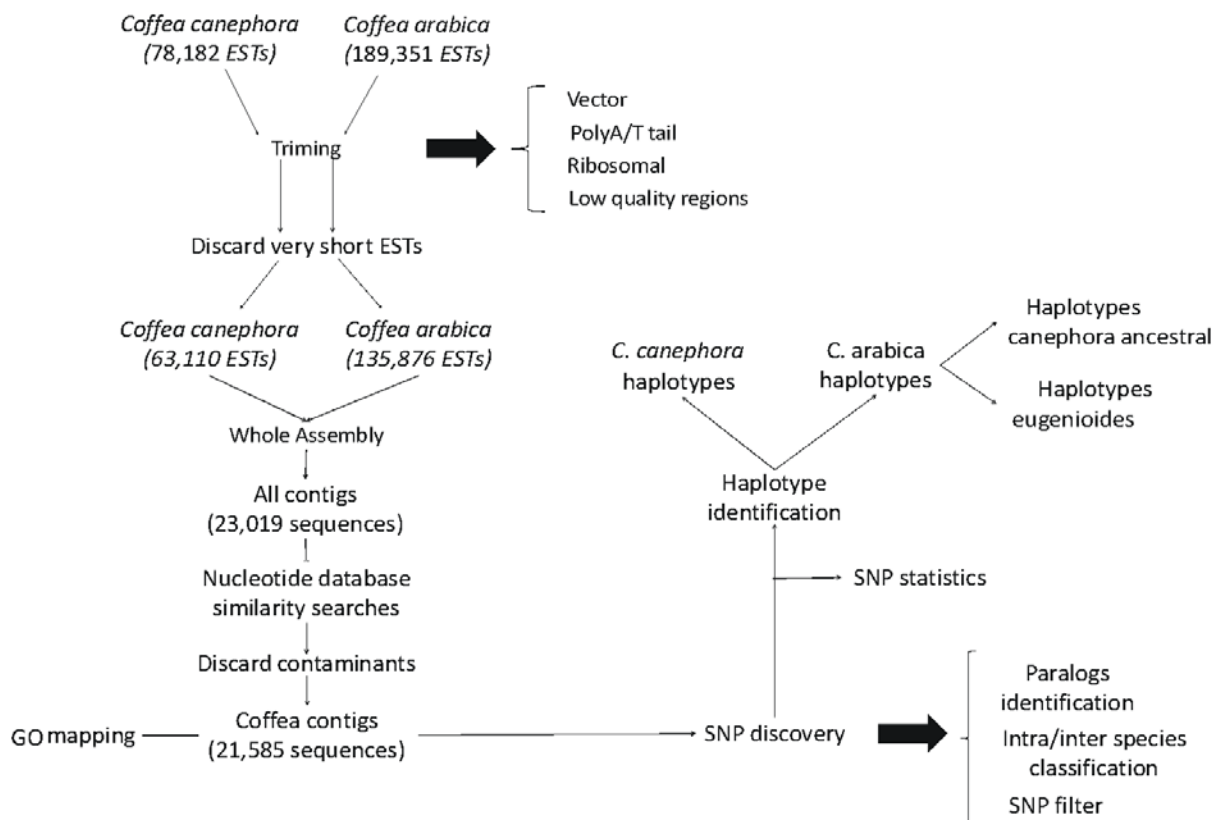


Figura 1. Pipeline utilizado para a montagem e descoberta de SNPs em larga escala.

Através do agrupamento dos haplótipos de *C. arabica* com *C. canephora* foi possível identificar que aproximadamente 45% de todos os ESTs expressos tem origem no ancestral canephora e 55% no ancestral eugenoides. Os genes cujo maior responsável pelos transcritos foi o *C. eugenoides* estão mais relacionados com metabolismo basal e metabolismo secundário, enquanto os genes com mais transcritos de origem de *C. canephora* estão mais envolvidos com transdução de sinal e regulação da expressão gênica.

Entre os dois genótipos de *C. arabica* disponíveis no banco de dados do Projeto Genoma Café Brasileiro apenas 13 polimorfismos foram encontrados em 1847 Contigs (unigenes). Devido essa baixa diversidade observada, foi iniciado um trabalho de sequenciamento de genes relacionados com qualidade em genótipos de *C. arabica* representativos do centro de origem, além de cultivares comerciais.

Foi realizada a amplificação dos 16 genótipos com 16 pares de primers, sendo 13 relacionados ao metabolismo de açúcares e diterpenos, 1 com ADH (álcool desidrogenase) e 2 com genes do cloroplasto e da mitocôndria - para estudo comparativo da evolução dos genes candidatos. Após verificar que todos os genótipos produziam fragmentos amplificados com todos os primers previamente definidos, as amostras foram amplificadas, purificadas e sequenciadas.

Até o presente momento somente as sequências obtidas para um fragmento (679 bp) do gene que codifica para SPS foram analisadas. Vinte e dois polimorfismos incluindo 21 SNP e um microssatélite foram detectados (Tabela 1). Na maioria dos casos os polimorfismos detectados correspondem à polimorfismos entre *Coffea canephora* e *Coffea eugenoides* (10 polimorfismos) e polimorfismos detectados entre genótipos de *Coffea canephora* (9 polimorfismos). Os

três últimos polimorfismos correspondem a polimorfismos entre *Coffea racemosa* e *Coffea arabica* e *Coffea canephora* e a polimorfismos heterozigotos na planta de *Coffea racemosa* analisada (Tabelas 1 e 2).

Tabela 1 : Características dos polimorfismos detectados num fragmento do gene SPS (679 bp) num painel de 21 genótipos de *Coffea* spp

Nome do polimorfismo	Sequência antes do polimorfismo	Sequência depois do polimorfismo	Tipo de polimorfismo	Allelo1	Allelo 2	Nível do polimorfismo
1	AAAAGGAATTTACAACACCTATAC	GGCATAACTTAGTTGGCCACTTTA	SNP	A	G	Entre Cc e Ce
2	AATTTACAACACCTATACGGCATAACTT	GTTGGCCACTTTAGTTTAA	SNP	A	C	C. r ou entre Cc e Cr
3	CTTAGTTGGCCACTTTAGTTTA	CCCATTTATCAACAGGTGCGAAAGCAC	SNP	A	G	Entre Cc e Ce
4	CAACCGTGCCAGTTTCAAAGC	AGAACGAAATTAACGAATAGAATCATAC	SNP	A	G	C. c
5	CCGTGCCAGTTTCAAAGCAAGAAC	AAATTAACGAATAGAATCATAC	SNP	C	G	C. c
6	ATAGAATCATACACTGATGG	TTTTATCAGGGTAT	SNP	C	T	C. c
7	TACACTGATGGTTTTTATCAGGG	ATCATAACAAGGTAGC	SNP	C	T	Entre Cc e Ce
8	GTTTTTATCAGGGTAT	ATAACAAGGTAGCAAACTTTT	SNP	C	T	C. c
9	CAGGCCAATTATTAACACGA	GTTTCATGTCAAACCTTTCTGT	SSR	(AT)3	(AT)4	Entre Cc e Ce
10	TCATGTCAAACCTTTCTGTG	TGTGCGGAAGTGCAAAATTTCTAGCTTTT	SNP	A	C	Entre Cc e Ce
11	TAGCTTTTGAGAACCCCTAGCTTTTC	AACGAGTCGCGCAGCTCAGCAC	SNP	C	T	Entre Cc e Ce
12	AGAACCCTAGCTTTTCCAA	GAGTCGCGCAGCTCAGCAC	SNP	C	T	Entre Cc e Ce
13	TGAGAACCCCTAGCTTTTCCAACGA	TCGCGCAGCTCAGCACTGCTG	SNP	C	G	C. c
14	AGCTTTTCCAACGAGTCGCGCAG	TCAGCACTGCTGCAGTCTTCACT	SNP	C	T	Entre Cc e Ce
15	TTCCAACGAGTCGCGCAGCTCAGC	CTGCTGCAGTCTTCACTTGTG	SNP	A	T	C. c
16	CAACGAGTCGCGCAGCTCAGCAC	GCTGCAGTCTTCACTTGTCTGAA	SNP	A	T	Entre Cc e Ce
17	TCCAACGAGTCGCGCAGCTCAGCACT	CTGCAGTCTTCACT	SNP	G	T	Entre Cc e Ce
18	AGCTCAGCACTGCTGCAGTCTTCACT	GTCTGAATAAGGTTCCGATTGTCA	SNP	C	T	C. c
19	TGCCACCACATCGGTAAG	GGGTAGCTTCTGTGGCATGAAGTTGG	SNP	C	T	C. r ou entre Cc e Cr
20	GTTGGCTGCTTTCTCCACTGCAAAAC	CCTTTCAGTATCAGCACT	SNP	C	T	C. c
21	CCTTTCAGTATCAGCACTT	TGCACACCACCAAGTAATCCTT	SNP	A	G	C. c
22	CTTTCTCCAACAAAACCAACCAC	TTTGACAAATCCATACCCCATCG	SNP	C	T	C. r ou entre Cc e Cr

Tabela 2- Diversidade nucleotídica do fragmento do gene SPS (679 bp) dentro dum painel de 21 genótipos de *Coffea* spp. Linhas correspondem aos genótipos. Colunas com números (1-22) correspondem ao polimorfismo encontrado de acordo com a tabela 1.

Genótipo	Espécie	Grupo genético <sup>a</sup>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
E-017	C. arabica	Ethiopian3	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-123A	C. arabica	Ethiopian4	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-123B	C. arabica	Ethiopian4	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-237	C. arabica	Ethiopian2	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-238	C. arabica	Ethiopian4	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-464	C. arabica	Ethiopian1	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E-516	C. arabica	Ethiopian1	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bourbon	C. arabica	Bourbon	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Typica	C. arabica	Typica	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Catuai	C. arabica	—	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Mundo Novo	C. arabica	—	A/G	A	A/G	A	G	T	C/T	C/T	3/4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G2020	C. canephora	Guinea	G	A	A	A	G	C	T	T	4	A	C	C	G	C	A	T	G	T	T	C	A	NA
C1007	C. canephora	Congolese_B	G	A	A	A	C	T	T	C	4	A	C	C	C	C	A	T	G	C	T	T	G	T
C4001	C. canephora	Congolese_C	G	A	A	A	C/G	T	T	C	4	A	C	C	G	C	A	T	G	C	T	T	A	NA
C3001	C. canephora	Congolese_SG1	G	A	A	A/G	G	T	T	C	4	A	C	C	C	C	A	T	G	C	T	T	A	NA
G21	C. canephora	Congolese_SG1	G	A	A	A/G/C/G	T	T	C	4	A	C	C	C	C	A/T	T	G	C	T	T	A/G	T	NA
C2011	C. canephora	Congolese_SG2	G	A	A	A	C	T	T	C	4	A	C	C	C/G	C	A	T	G	C	T	T	A/G	NA
UW099	C. canephora	Uganda_Selvagem	G	A	A	A	C	T	T	C	4	A	C	C	C/G	C	A	T	G	C/T	T	T	A/G	NA
UW002	C. canephora	Uganda_Selvagem	G	A	A	A	C	T	T	C	4	A	C	C	G	C	A	T	G	T	T	T	A	NA
Coffea eugenioides	C. eugenioides	—	A	A	G	A	G	T	C	T	3	C	T	T	G	T	A	A	T	T	T	T	A	T
Coffea racemosa	C. racemosa	—	A	A/C	A	A	G	T	T	T	3	A	T	C	G	T	A	T	G	T	C/T	T	A	C

<sup>a</sup> Grupos genéticos baseados em Anthony et al (2001) para *Coffea arabica* e Montagnon (2000) para *C. canephora*

NA : informação não disponível, apos o polimorfismo 9 (microsatélite), heterozigoto nos acessos de *C. arabica* não foi possível analisar as sequências.

Em relação aos 11 genótipos de *Coffea arabica* sequenciados, 5 polimorfismos foram detectados em uma região de 295 bases. Todos esses polimorfismos estão num estado heterozigoto e correspondem à polimorfismos interspecíficos entre *C. canephora* e *C. eugenioides* (Tabela 2). A presença de um “motif” microsatélite heterozigoto (polimorfismo 9 na Tabela 1) em todos os genótipos de *Arabica* não permitiu a análise das sequências após esse polimorfismo. Apesar da utilização de genótipos dos diferentes grupos de diversidade de *Coffea arabica* (Anthony et al., 2001) nenhum polimorfismo foi detectado entre os genótipos de *Arábica* analisados. Isso pode ser explicado pelo fato da *Coffea arabica* ser uma espécie aloploplóide autógama recente o que contribui para uma variabilidade genética baixa. Além disso, nesse ensaio preliminar, as sequências foram feitas sem clonagem dos fragmentos, o que dificulta a detecção de polimorfismos presente em baixas doses, ou seja, um alelo mutante dentro dos quatro possíveis. As sequências para os demais genes estão sendo analisadas para confirmar os resultados obtidos para o fragmento de SPS.

## CONCLUSÕES

Os resultados obtidos, baseados no sequenciamento direto de produto PCR (sem clonagem), sugerem que para o pequeno número de genes analisados até agora, a diversidade nucleotídica presente dentro de cada genoma ancestral de *Coffea arabica* é limitada. Os resultados também indicam a necessidade da clonagem do produto de PCR para *C. arabica* antes do sequenciamento para permitir a análise da diversidade de cada homólogo com uma melhor precisão no caso da existência de polimorfismos presentes em baixas doses.

## AGRADECIMENTOS

Esse projeto foi financiado pelo CBPD- Café. SD Lannes e LP Ferreira –Bolsistas do CBPD-Café. K Yanagui bolsista do Programa de Iniciação Científica do IAPAR – ProICI. Ramon Vidal Bolsista FAPESP.

## REFERÊNCIAS BIBLIOGRÁFICAS

- Anthony F et al. (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. *Euphytica* 118: 53–65
- Lin C. et al. 2005. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts *Theor Appl Genet.* 112(1): 114–130
- Montagnon C (2000). Optimisation des gains génétiques dans le schéma de sélection récurrente réciproque de *Coffea canephora* Pierre. ENSA Montpellier, France, PhD thesis
- Risterucci AM et al. (2000). A high-density linkage map of *Theobroma cacao* L. *Theor Appl Genet.* 101(5-6): 948-955
- Vieira et al. 2006. Brazilian coffee genome project: an EST-based genomic resource *Braz. J. Plant Physiol.*, 18(1):95-108